BAB I
PENDAHULUAN

Bab ini menjelaskan tentang latar belakang permasalahan yang mendasari pembuatan Tesis, batasan masalah, rumusan masalah, serta tujuan dan manfaat penelitian.

1.1 Latar Belakang Penelitian

Perkembangan teknologi digital telah membawa perubahan signifikan dalam cara masyarakat berinteraksi dan menyampaikan opini. Twitter, yang kini dikenal sebagai X, menjadi salah satu platform media sosial dengan pengguna aktif terbesar di Indonesia, dan banyak dimanfaatkan untuk mengekspresikan pendapat, sentimen, serta reaksi terhadap berbagai isu publik secara *real-time*. Data sentimen yang dihasilkan dari platform ini memiliki nilai strategis bagi pengambil keputusan, baik dalam pemantauan reputasi merek, evaluasi kebijakan pemerintah, maupun pemahaman terhadap opini publik secara lebih luas.

Namun, bahasa yang digunakan pengguna Twitter Indonesia bersifat sangat informal dan dinamis. *Tweet* umumnya memuat bahasa gaul (*slang*), singkatan tidak baku, ejaan yang tidak standar, karakter berulang, penggunaan emoji, serta pencampuran bahasa Indonesia dengan bahasa Inggris atau bahasa lainnya (*code-mixing*). Karakteristik ini menjadikan analisis sentimen pada teks Twitter berbahasa Indonesia jauh lebih sulit dibandingkan pada teks formal, dan menuntut pendekatan pemrosesan bahasa alami (*Natural Language Processing*) yang mampu menangkap isi teks secara lebih mendalam dibandingkan metode konvensional.

Kemunculan model bahasa pra-latih berbasis *Transformer*, seperti mBERT (*Multilingual* BERT) dan IndoBERT (*Indonesian* BERT), membuka peluang baru untuk meningkatkan performa analisis sentimen berbahasa Indonesia. Kedua model ini memiliki karakter yang berbeda: mBERT dilatih pada korpus dari 104 bahasa sekaligus, sehingga kemampuan *cross-lingual* yang dimilikinya

luas namun distribusi kapasitas representasinya juga terbagi ke banyak bahasa. Sebaliknya, IndoBERT dilatih khusus pada korpus Indo4B yang memuat lebih dari 4 miliar kata berbahasa Indonesia dari berbagai sumber, termasuk media sosial dan dokumen informal. Perbedaan filosofi pelatihan ini memunculkan pertanyaan mendasar, yang mana model pra-latih khusus untuk satu bahasa apakah benar-benar lebih unggul dibandingkan model multibahasa ketika diterapkan pada teks Twitter Indonesia yang sarat bahasa nonbaku dan campuran bahasa (*code-mixing*).

Di sisi lain, validitas hasil analisis sentimen juga sangat bergantung pada kualitas dan karakteristik dataset yang digunakan. Dataset yang baik perlu relevan dengan topik yang dikaji, merepresentasikan karakteristik data secara memadai, serta memiliki distribusi kelas sentimen yang dapat diidentifikasi dengan jelas, mengingat ketimpangan proporsi antar kelas dapat memengaruhi kemampuan model dalam melakukan klasifikasi secara akurat. Pertimbangan inilah yang mendasari perlunya pemahaman menyeluruh terhadap karakteristik dataset sebelum menilai performa suatu model secara adil.

Penelitian ini menggunakan *Indonesian General Sentiment Analysis Datase*t, yang telah divalidasi sebagai *dataset* yang dapat menghasilkan nilai akurasi setara dengan benchmark internasional SemEval (Ferdiana dkk., 2019). Dataset tersebut berisi 10.806 *tweet* dengan tiga kelas sentimen, dan pada penelitian sebelumnya oleh Ferdiana telah diuji menggunakan metode klasik dengan akurasi tertinggi sebesar 62,9% menggunakan *classifier Support Vector Machine* (*SVM*). Angka tersebut mengindikasikan masih besarnya ruang perbaikan. Meskipun mBERT dan IndoBERT menawarkan potensi peningkatan performa, hingga saat ini belum ditemukan penelitian yang membandingkan keduanya secara langsung pada dataset yang sama. Selain belum adanya perbandingan langsung, terdapat pula kesenjangan metodologis: penelitian-penelitian terdahulu umumnya mengutip angka *accuracy* dari literatur sebagai pembanding tanpa mereplikasi *baseline* pada kondisi identik (kesenjangan replikasi), tidak menyertakan audit sistematis terhadap proses tokenisasi sebagai penjelas perbedaan performa (kesenjangan metodologis pada

level analisis), dan jarang memvalidasi perbedaan performa antar model secara statistik (kesenjangan validitas). Penelitian ini juga mengisi kesenjangan kontekstual, karena mayoritas studi pembanding IndoBERT-mBERT dilakukan pada domain non-Twitter atau tanpa penekanan eksplisit pada karakteristik *code-mixing* dan bahasa informal yang menjadi ciri khas platform.

Penelitian ini hadir untuk mengisi kesenjangan tersebut dengan membandingkan mBERT dan IndoBERT secara empiris pada tugas klasifikasi sentimen tiga kelas, disertai audit komprehensif terhadap proses tokenisasi kedua model serta replikasi baseline TF-IDF dan *SVM* pada *test* set yang identik. Melalui pendekatan ini, penelitian diharapkan mampu menghasilkan rekomendasi pemilihan model yang dapat dipertanggungjawabkan secara empiris maupun metodologis bagi konteks analisis sentimen Twitter berbahasa Indonesia yang berkarakteristik informal dan *code-mixing*.

1.2 Batasan Masalah

Berdasarkan permasalahan yang disebutkan di atas, batasan masalah pada penelitian ini adalah sebagai berikut.

1. *Dataset* yang digunakan adalah *Indonesian General Sentiment Analysis Dataset* yang berisi 10.806 *tweet* berbahasa Indonesia yang dikumpulkan dari media sosial Twitter.
2. Penelitian ini berfokus pada klasifikasi sentimen pada tingkat kalimat (*sentence-level sentiment classification*) dengan tiga kategori sentimen, yaitu positif, negatif, dan netral.
3. Model yang digunakan adalah *Multilingual* BERT versi bert-*base-multilingual-uncased* dan IndoBERT versi *indobenchmark/indobert-base-p1* (varian *base*, 12 *transformer* layers, 768 hidden units, ~110 juta parameter). Pendekatan yang digunakan adalah *fine-tuning* pada model *pre-trained* dengan *classifier* head tambahan untuk klasifikasi tiga kelas.
4. Distribusi kelas pada *dataset* diperiksa, dan apabila ditemukan ketidakseimbangan, diterapkan teknik *class weighting* menggunakan

rumus *balanced class weight* pada *loss function CrossEntropy*. Bobot tiap kelas dihitung berdasarkan formula n / (k × n_kelas), dengan n adalah jumlah seluruh sampel, k adalah jumlah kelas, dan n_kelas adalah jumlah sampel pada kelas tertentu.

5. Evaluasi performa model dilakukan menggunakan metrik *accuracy, precision, recall, dan F1-Score (macro, weighted)* untuk mengukur kemampuan model dalam mengklasifikasikan sentimen pada *dataset*. Selain evaluasi kuantitatif tersebut, penelitian ini juga melakukan analisis kesalahan (*error analysis*) untuk mengidentifikasi pola kesalahan klasifikasi yang terjadi pada masing-masing model. Analisis juga dilakukan terhadap *confusion matrix* untuk melihat kecenderungan kesalahan antar kelas sentimen.
6. Analisis terhadap proses tokenisasi pada kedua model, khususnya pada kata tidak baku, *slang*, serta kata yang mengandung pencampuran bahasa yang umum ditemukan pada teks media sosial guna memahami bagaimana perbedaan segmentasi *token* dapat mempengaruhi representasi teks dan hasil klasifikasi sentimen.
7. Penelitian ini juga menjalankan *baseline* klasik *TF-IDF* dengan klasifier *Linear Support vector machine* sebagai pembanding untuk mengetahui peningkatan nyata yang diberikan oleh pendekatan berbasis *transformer*. Konfigurasi *TF-IDF + SVM* dicari yang terbaik melalui metode *Grid search Cross validation*.

1.3 Rumusan Masalah

Berdasarkan uraian latar belakang, rumusan masalah dari penelitian ini adalah:

1. Bagaimana karakteristik distribusi label dan bahasa informal pada dataset yang digunakan, serta pengaruhnya terhadap pemilihan metodologis dalam penelitian?

2. Seperti apa hasil proses tokenisasi antara mBERT dan IndoBERT terhadap kata tidak baku, *slang*, singkatan, serta kata yang mengandung pencampuran bahasa dalam *dataset*?
3. Seberapa besar perbedaan performa model mBERT dan IndoBERT dalam melakukan klasifikasi sentimen pada *dataset tweet* berbahasa Indonesia berdasarkan metrik *accuracy, precision, recall*, dan *F1-Score* (*macro* dan *weighted*), serta perbandingannya terhadap *baseline TF-IDF SVM*?
4. Apa pengaruh karakteristik *dataset* dan perbedaan proses tokenisasi terhadap performa klasifikasi sentimen yang dihasilkan oleh kedua model?

1.4 Tujuan Penelitian

Tujuan penelitian ini disusun untuk menjawab rumusan masalah yang telah ditetapkan, yaitu sebagai berikut.

1. Mengidentifikasi karakteristik *Indonesian General Sentiment Analysis Dataset*, termasuk distribusi label sentimen serta penggunaan bahasa informal dalam *tweet*.
2. Menganalisis perbedaan proses tokenisasi pada mBERT dan IndoBERT terhadap kata tidak baku, *slang*, singkatan, serta kata yang mengandung pencampuran bahasa dalam *dataset*.
3. Mengukur dan membandingkan performa model mBERT dan IndoBERT dalam melakukan klasifikasi sentimen pada *dataset* menggunakan metrik accuracy, precision, recall, dan *F1-Score* (macro), serta perbandingan terhadap *baseline* TF-IDF *SVM*.
4. Menganalisis pengaruh karakteristik *dataset* dan perbedaan proses tokenisasi terhadap performa klasifikasi sentimen yang dihasilkan oleh mBERT dan IndoBERT.

1.5 Manfaat Penelitian

1.5.1 Manfaat Akademis

Penelitian ini diharapkan memberi manfaat akademis seperti penambahan wawasan tentang penggunaan mBERT dan IndoBERT dalam analisis sentimen Twitter berbahasa Indonesia, bukti empiris kontribusi pra-pelatihan domain melalui perbandingan terhadap *baseline* klasik *TF-IDF + SVM*, serta metodologi audit *tokenizer* komprehensif yang dapat direplikasi pada penelitian NLP Bahasa Indonesia lainnya.

1.5.2 Manfaat Praktis

Manfaat praktis yang diharapkan didapat dari penelitian ini yaitu memberikan gambaran umum mengenai pengolahan sentimen masyarakat sesuai *dataset* Twitter, memberikan gambaran tingkat *accuracy* dari metode mBERT dan IndoBERT untuk dijadikan acuan pengembangan penelitian berikutnya, serta memberikan rekomendasi kombinasi *hyperparameter* (*batch size, learning rate, epoch*) yang menghasilkan performa optimal untuk *fine-tuning* model BERT pada teks Twitter berbahasa Indonesia, sehingga dapat menjadi acuan praktis bagi peneliti dan praktisi yang membangun sistem analisis sentimen serupa.